Homework 2 COMP5423

NetID: 24133101g WU Yifan

Question 1:

Yes. Here is my detail answer.

Labelling task

For every character in a Chinese sentence will be labelled as B (beginning of a word), M(middle part of a word), E(end of a word) or S(single Chinese word). For example, “我喜欢苹果” will be labelled as "我/S 喜/B 欢/E 苹/B 果/E".

Designing part

Chinese semantic understanding is always bidirectional due to the contextual dependencies, segmentation can occur frequently (e.g. “研究生命”). I want to use a bidirectional LSTM to scan character sequences from both sides. To ensure segmentation accuracy, CRF layer can also be added to learn the label transition rules, such as B can only be followed by M or E, not S.

Architecture of RNN

Embedding layer: convert each Chinese character into a dense vector representation.

biLSTM layer: output feature vectors for each character based on the bidirectional contextual.

CRF layer: learn label transition rules.

During training, I may use a combined loss function of cross-entropy loss and CRF negative log-likelihood loss function.

BERT encoder

For pre-trained model, I may add a fully connected layer on top of each BERT token output and use softmax in the final layer to do a label classification. Optionally, add CRF layer to learn rules is also considered. Use the same loss function as RNN model.

Question 2:

Yes. N-gram combined Conditional Random Fields (CRF) or Hidden Markov Models (HMM) an also be applied to the Chinese word segmentation. I think about that the input features may conclude:

- n-gram combination features
- Lexicon features (presence in corpus and word frequency)
- Traditional statistical features (character occurrence frequency)
- Contextual features (windows-based contextual information)

In the features design, we can also consider the hybrid features or using dynamic weight features to improve the performance of the non-neuro model. For example, we can extract many features, however, use Adaboost or Gradient Boosting to automatically assign feature importance weights instead of manual settings.

Question 3:

Yes. Considering the encoder-decoder are frequently used to sequence-to-sequence (Seq2Seq) tasks, Chinese word segmentation can also be solved with this method. For example, for the input Chinese sequence “我喜欢苹果” can generate the output sequence “我 喜欢 苹果”. I design a simple encoder-decoder architecture to achieve it

Encoder: Same as traditional sequence labelling models (biLSTM or Transformer).

Decoder: for this decoder, it should have multiple tasks. The basic task is to predict the labels B, M, S, E. It is a traditional classification task. And the core task is to generate segmented result with delimiters based on the classification result.

Training Data: The both label sequences and segmented results. The loss function is the cross-entropy loss (including sequence classification cross-entropy loss and encoder-decoder cross-entropy loss)

However, compared to other methods, the encoder-decoder architecture is less efficient, requires higher quality data, and the model is more complex and the error rate is relatively high.